Le Coran : quelques données lexico-statistiques[1]

Djamel Eddine Kouloughli[2]

0. Préambule

Comme les textes sacrés des autres religions monothéistes, le Coran a fait l'objet, très tôt, et en tout cas bien avant l'invention de l'ordinateur, de décomptes statistiques minutieux portant sur tous les aspects de son organisation textuelle. C'est ce qu'atteste par exemple un célèbre récit rapporté dans *Les Mille et une Nuits* et qui met en scène « la docte Sympathie », belle et savante jeune femme qu'un aréopage de savants interroge, notamment sur le Coran. Voici un extrait de sa réponse :

> Elle répondit : « Le Korân est composé de cent quatorze sourates ou chapitres, dont soixante-dix ont été dictés à La Mecque et quarante-quatre à Médine. Il est divisé en six cent vingt-et-une divisions, appelées "aschar", et en six mille deux cent trente-six versets. Il renferme soixante dix-neuf mille quatre cent trente-neuf mots, et trois cent vingt-trois mille six cent soixante-dix lettres »[3]...

Nos propres données, et nos moyens actuels de les traiter, ne confirment pas cette description dans ses moindres détails. Mais il faut bien reconnaître que tous les ordres de grandeur évoqués dans ce texte sont corroborés par nos propres résultats, ce qui, compte tenu du fait que tous les décomptes étaient faits à la main, ne peut que susciter l'admiration[4].

Aujourd'hui, des textes électroniques du Coran sont très faciles à trouver et à télécharger sur le web et, moyennant quelques manipulations, à transformer en corpus électroniques susceptibles d'être analysés de façon exhaustive et systématique. Ce sont essentiellement des résultats de telles analyses qui seront présentés et commentés dans les pages qui suivent.

Le corpus sur lequel nous avons travaillé, comme la plupart des versions électroniques du Coran en circulation, correspond à ce que l'on désigne parfois sous le nom de « Coran de Fouad ». Cette dénomination fait référence à la pre-

---

1. Le présent texte corrige une version précédente où des erreurs s'étaient glissées dans les données statistiques. Ces erreurs nous ont été signalées par M. Tidjani Négadi, du département de physique de l'université d'Oran (Algérie). Qu'il trouve ici l'expression de nos remerciements.

2. CNRS, HTL (Histoire des théories linguistiques), UMR 7597.

3. « Histoire de la docte Sympathie » (178e nuit). *Les Mille et une Nuits*, traduction Joseph Charles Mardrus.

4. Surtout si l'on tient compte du fait qu'une partie des divergences vient peut-être de la définition des observables (par exemple pour ce qui concerne les « mots » et les « lettres »).

mière version imprimée du Coran, laquelle a été élaborée sous l'autorité des *ʿulamāʾ* d'al-ʾAzhar et sous l'égide du roi Fouad Ier d'Égypte en 1923. Elle correspond à la « lecture » (*qirāʾa*) de Ḥafṣ telle que transmise par ʿĀṣim, traditionnellement en usage dans le pays du Nil, et devenue, par les vertus démocratiques de l'imprimerie, la version écrite du Coran la plus largement diffusée dans le monde.

Le Coran relève, à l'origine, de l'oralité, comme le suggère la probable étymologie de *Qurʾān*, « récitation » voire « appel, proclamation ». De la relativement longue période où il a été essentiellement confié à la transmission orale assurée par les « porteurs du Coran » (*ḥamalat al-Qurʾān*), il a gardé des variantes textuelles plus ou moins importantes auxquelles la tradition fait référence sous le nom de *ʾaḥruf* (idiomes), d'une part, de *qirāʾāt* (récitations, lectures), d'autre part. Les *ʾaḥruf*, dont une tradition nous apprend que le Prophète aurait lui-même déclaré que la révélation en comportait sept, semblent concerner des différences relatives à l'identité de certaines unités lexicales du texte qui se manifestent dans le ductus consonantique (*rasm*). Les *qirāʾāt* concernent plutôt des différences dans les marques diacritiques associées à ce ductus (vocalisation et gémination consonantique). La fixation graphique définitive du Coran, qui ne se fait que progressivement, avec l'adoption généralisée de la réforme de l'écriture arabe décidée par le Calife omeyyade ʿAbd al-Malik Ibn Marwān (m. 705), s'accompagnera d'un net recul de la tolérance, assez grande à l'origine, à l'égard des variantes dans la récitation du texte sacré. Au Xe siècle, à l'initiative du théologien Ibn Muǧāhid (m. 936), seules sept *qirāʾāt* « canoniques » sont codifiées et retenues pour être reproduites. Les autres, même si elles sont notées, sont déclarées « rares » (*šāḏḏa*) et pratiquement ostracisées[5]. Des sept *qirāʾāt* reconnues, seules deux ont eu une diffusion assez large dans la communauté musulmane sunnite, celle de Ḥafṣ dont nous avons déjà parlé, et celle de Warš, encore en usage au Maghreb. Le « Coran de Fouad » semble avoir fait définitivement pencher la balance en faveur de la première.

1. Présentation générale

Dans sa version électronique comme dans sa version imprimée, le texte coranique est divisé en 114 « sourates » (*sūra/suwar*[6]) classées dans un ordre conventionnel (c'est-à-dire non chronologique). Chaque sourate porte un ou plusieurs noms, également conventionnels[7]. Concernant le classement des sourates du

5. Sur l'histoire de la fixation du texte coranique on peut lire, dans les références arabes anciennes, l'ouvrage de ʾAbū ʿAmr al-Dānī, *Al-Muqniʿ fī rasmi maṣāḥifi al-ʾamṣār*, et dans les références modernes, par exemple Blachère (1958) ou Déroche (2005).

6. Sur l'étymologie du mot *sūra*, il existe, tant en arabe que dans diverses langues européennes, une abondante littérature, abondance qui témoigne surtout de l'absence d'explication vraiment convaincante de la signification originelle du mot. Tout au plus peut-on dire, avec certitude, que la racine <*swr*> à laquelle est associé ce mot porte le contenu sémantique général de « barrière, séparation ».

7. Ainsi, la première sourate du Coran, outre son nom de *Fātiḥa* (« Introduction »), est connue sous plus d'une dizaine d'autres dénominations. De même, la *sūra* 112, à laquelle il

Coran, on dit souvent qu'à l'exception de la première, la *Fātiḥa*, elles seraient classées par ordre de longueur décroissante. Cela n'est que tendanciellement vrai. Dans le détail, cette description ne tient pas : par exemple, la sourate 7 (*al-'Aʿrāf*) compte 204 versets alors que les trois sourates qui la précèdent en comptent moins de 200. Elle est en outre suivie de la sourate 8 (*al-'Anfāl)* qui ne compte que 76 versets, alors que les quatre sourates suivantes (9, 10, 11 et 12) comptent toutes plus de 100 versets. On pourrait multiplier les exemples de ce genre, qui montrent que, *stricto sensu*, il est inexact de dire que les sourates du Coran sont classées par ordre de longueur décroissante.

Cela soulève, soit dit en passant, une question qui sort de l'objet de la présente communication, mais qui a son importance, celle de savoir quel est le principe qui a présidé au classement des sourates du Coran tel qu'il nous est parvenu. Deux hypothèses au moins se proposent : la première, assez superficielle, est qu'il s'agit d'un classement visant à optimiser les stratégies de mémorisation du texte. La seconde, évoquée notamment par le grand compilateur de la culture arabe traditionnelle que fut al-Suyūṭī (m. 1505), dans *'Asrār tartīb al-Qur'ān* (« Les secrets du classement du Coran »), est que l'analyse thématique interne des sourates révélerait un fil conducteur (voire plusieurs) conduisant d'une sourate à l'autre dans leur ordre de classement tel qu'il est attesté. Cette dernière hypothèse mériterait d'être explorée systématiquement.

En en-tête de chaque sourate on trouve, parfois, dans les versions électroniques, et toujours dans les versions imprimées, une indication relative à l'origine, mecquoise ou médinoise, de la sourate. Rappelons à ce propos que le Coran a été révélé au Prophète de l'islam sur une durée de quelque 23 ans, et qu'une partie de cette révélation a été reçue pendant qu'il résidait à La Mecque, sa ville natale, et l'autre à Médine où, à partir de 622, il émigre avec ses partisans pour jeter les bases du futur État islamique. C'est sur cette périodisation que se fonde l'indication de l'origine, mecquoise ou médinoise, des sourates. Cette indication, qui donne une première chronologie globale des sourates composant le texte coranique, soulève bien sûr la question du classement chronologique détaillé des sourates, c'est-à-dire celle de l'ordre de succession exact des unes par rapport aux autres. Certains seront peut-être surpris d'apprendre que sur cette question essentielle[8], il n'y a pas vraiment de *consensus doctorum*.

Par exemple, l'un des textes les plus anciens relatifs à la chronologie coranique, le *Kitābu tanzīli l-Qur'ān* (« Livre de la révélation du Coran ») d'Ibn Šihāb al-Zuhrī (m. 741) indique que la *Fātiḥa* aurait été révélée à Médine alors que presque tous les auteurs postérieurs affirment qu'il s'agit d'une sourate mecquoise. Selon

---

est généralement référé comme *sūrat al-'Iḵlāṣ*, est connue sous une douzaine d'autres noms.

8. Cette question est essentielle, notamment d'un point de vue juridique, dans la mesure où certaines prescriptions coraniques ont été modifiées au cours de la révélation. Il importe donc de savoir laquelle des deux prescriptions différentes relatives au même domaine (par exemple la consommation de vin) est chronologiquement la dernière et donc celle dont le respect s'impose.

al-Zuhrī, le nombre des sourates révélées à La Mecque est de 85 et celui des sourates médinoises de 29, chiffres à comparer avec ceux que donnait « la docte Sympathie » dans le texte des *Mille et une Nuits* (*supra*) et qui reflète une position traditionnelle courante. À titre indicatif, voici la classification chronologique des sourates selon al-Zuhrī[9] (le premier chiffre renvoie à la classification conventionnelle, le second (après le slash) à l'ordre chronologique selon al-Zuhrī) :

Sourate mecquoises :
96/1;68/2;73/3;74/4;111/5;81/6;87/7;92/8;89/9;93/10;94/11;100/12;103/13;108/14;
102/15;107/16;109/17;105/18;113/19;114/20;112/21;53/22;80/23;97/24;91/25;85/26;
95/27;106/28;101/29;75/30;77/31;50/32;104/33;54/34;90/35;86/36;38/37;7/38;72/39;
36/40;25/41;35/42;19/43;20/44;56/45;26/46;27/47;28/48;17/49;10/50;11/51;12/52;
15/53;6/54;37/55;31/56;34/57;39/58;40/59;41/60;42/61;43/62;44/63;45/64;46/65;
51/66;88/67;18/68;16/69;71/70;14/71;21/72;23/73;32/74;52/75;67/76;69/77;70/78;
78/79;79/80;82/81;84/82;30/83;29/84;83/85;

Sourates médinoises :
1/86;2/87;8/88;3/89;33/90;60/91;4/92;99/93;57/94;47/95;13/96;55/97;76/98;65/99;
98/100;59/101;110/102;24/103;22/104;63/105;58/106;49/107;66/108;62/109;64/110;
61/111;48/112;5/113;9/114.

Signalons qu'il existe une classification chronologique « officielle » d'al-'Azhar, laquelle ne semble différer de celle d'al-Zuhrī que par la position accordée à la *Fātiḥa*, considérée comme mecquoise et cinquième sourate révélée. Dans le cadre de cette classification, que nous adopterons dans les décomptes présentés plus loin, il y a 86 sourates mecquoises et 28 sourates médinoises.

Certains arabisants occidentaux ont également proposé des classements chronologiques des sourates du Coran en se fondant en partie sur les indications fournies par les sources musulmanes traditionnelles et en partie sur la critique textuelle interne du Coran. La tentative la plus marquante reste celle de Nöldeke poursuivie par Schwally. Voici, selon les mêmes conventions de présentation que ci-dessus, la classification chronologique proposée par Nöldeke[10] :

Sourate mecquoises :
96/1;74/2;111/3;106/4;108/5;104/6;107/7;102/8;105/9;92/10;90/11;94/12;93/13;97/14;86/15;
91/16;80/17;68/18;87/19;95/20;103/21;85/22;73/23;101/24;99/25;82/26;81/27;53/28;84/29;
100/30;79/31;77/32;78/33;88/34;89/35;75/36;83/37;69/38;51/39;52/40;56/41;70/42;55/43;112/44;10
9/45;113/46;114/47;1/48;54/49;37/50;71/51;76/52;44/53;50/54;20/55;26/56;15/57;19/58;38/59;36/6
0;43/61;72/62;67/63;23/64;21/65;25/66;17/67;27/68;18/69;32/70;41/71;45/72;16/73;30/74;
11/75;14/76;12/77;40/78;28/79;39/80;29/81;31/82;42/84;10/84;34/85;35/86;7/87;46/88;6/89;
13/90;

---

9. On pourra vérifier que, selon al-Zuhrī, la Fātiḥa est la 86e sourate révélée, et la première révélée à Médine. Signalons que, dans le texte publié d'al-Zuhrī auquel nous avons eu accès, une erreur rend la classification chronologique inconsistante. Cette erreur réside dans le fait que la sourate 7 (*al-'A'rāf*) est classée en deux positions chronologiques différentes, car référencée sous deux noms différents, « *'alif-lām-mīm-ṣād* » d'abord puis *al-'A'rāf*. Le reclassement systématique que nous avons effectué montre qu'au lieu de *al-'A'rāf*, dans la deuxième position, il faut lire *al-'Aḥzāb* (sourate 33), moyennant quoi tout rentre dans l'ordre.

10. On notera que pour Nöldeke, la *Fātiḥa* est bien mecquoise (48e sourate mecquoise), et qu'il n'y a que 24 sourates médinoises.

Sourates médinoises :
2/91;98/92;64/93;62/94;8/95;47/96;3/97;61/98;57/99;4/100;65/101;59/102;33/103;63/104;24/105;58/106;22/107;48/108;66/109;60/110;110/111;49/112;9/113;5/114;

La distinction entre sourates mecquoises et sourates médinoises a des manifestations textuelles que les analystes du texte coranique, anciens ou modernes, n'ont pas manqué de souligner. La prose rimée (*saǧʿ*) caractérise l'ensemble du texte coranique et se manifeste par le fait que les formes pausales des fins de versets d'une même sourate riment généralement entre eux[11]. Mais ce phénomène s'exprime de façon beaucoup plus discrète dans les sourates médinoises que dans les sourates mecquoises, notamment en raison du fait que les sourates de la période médinoise sont en général plus longues que celles de la période mecquoise, et qu'en outre les versets y sont sensiblement plus longs. Les deux types de sourates présentent également un aspect stylistique et un contenu thématique sensiblement différent :

– les sourates mecquoises, en particulier les premières, ont souvent un rythme rapide, saccadé, et sont riches en performatifs (serments, menaces, interpellations). Avec le temps, le style en devient plus ample, plus imagé, plus lyrique, mais les effets de rythme et de rime y restent toujours sensibles. Globalement, le contenu thématique des sourates mecquoises est nettement orienté vers des exhortations à l'adoption de la nouvelle foi, accompagnées d'évocations eschatologiques ou de références à des figures bibliques exemplaires ;

– les sourates médinoises présentent des versets beaucoup plus amples, si bien que les effets de la rime, toujours présente en fin de verset, y sont nettement plus discrets. Le style y devient plus démonstratif et l'argumentation plus articulée. C'est dans les sourates de cette période que l'on trouve les développements doctrinaux, juridiques et réglementaires qui vont servir de base à l'organisation de la société musulmane. On y trouve aussi, cependant, de longs récits édifiants reprenant et réinterprétant certains grands thèmes des traditions monothéistes antérieures.

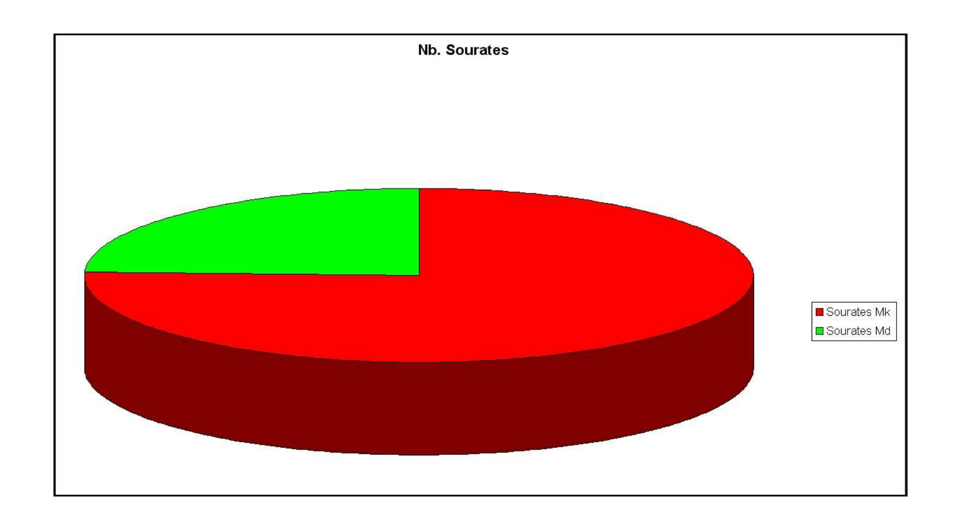
Au total, on se trouve, avec la structure générale du texte coranique dont on vient d'esquisser la présentation, et notamment avec la dichotomie sourates mecquoises/sourates médinoises, devant une dualité textuelle assez clairement marquée, mais dont les effets statistiques sont, d'une certaine manière, « neutralisés » : les sourates mecquoises, généralement moins longues mais aussi plus nombreuses que les sourates médinoises, l'emportent finalement en « couverture textuelle ». Les données statistiques suivantes illustrent cet état de choses :

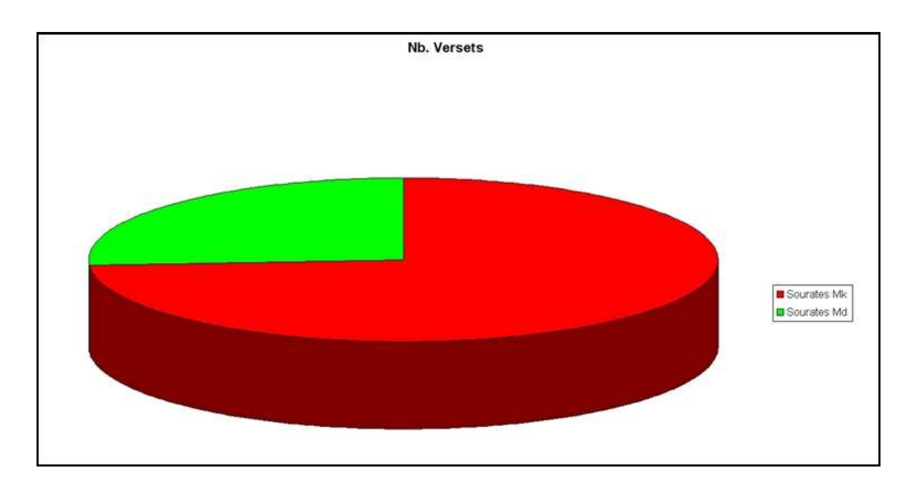
11. Il convient de noter que le type de rime dont fait usage le texte coranique, et que l'on appelle *fāṣila*, se réalise sur les formes pausales des mots et est donc nettement différent de la rime de la poésie arabe classique, dite *qāfiya*, qui se réalise sur les formes non pausales des mots.

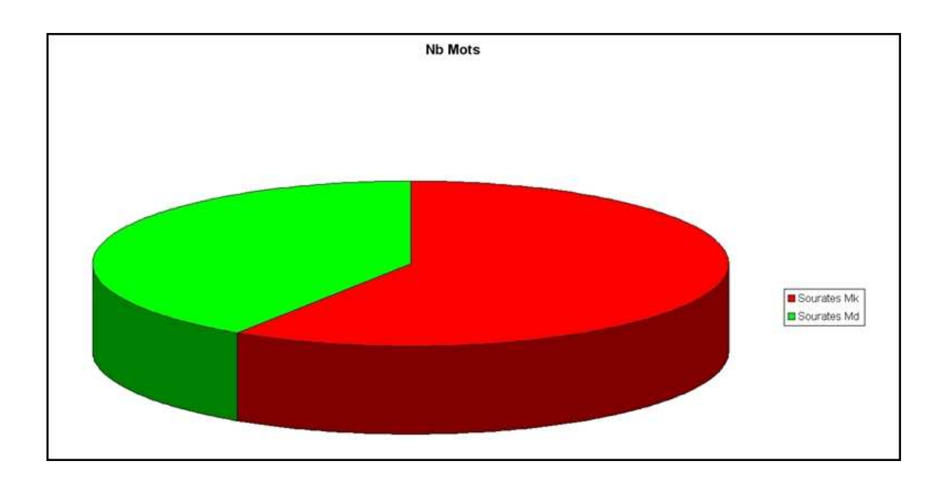
Tableau des sourates mecquoises et médinoises

	Nb. sourates	% sourates	Nb. versets	% versets	Nb. mots	% mots
Sourates Mk	86	75 %	4 613	74 %	47 692	59 %
Sourates Md	28	25 %	1 623	26 %	32 854	41 %
TOTAL	114	100 %	6 236	100 %	80 546	100 %

Dans le tableau ci-dessus comme dans les graphiques qui suivent, on a adopté les abréviations suivantes : Nb = nombre ; Mk = mecquoises ; Md = médinoises.







Pour clore le chapitre des généralités sur le texte coranique et plus spécifiquement sur la version électronique que nous avons soumise à l'analyse, précisons que, par rapport à certaines versions imprimées, cette version présente deux différences qu'il convient de souligner :

– d'une part, son orthographe est normalisée et « modernisée », c'est-à-dire qu'elle ne reproduit pas les variations orthographiques idiosyncrasiques que certaines versions imprimées conservatrices exhibent et qui font que le même mot, par exemple le nom propre *'Ibrāhīm*, peut apparaître sous plusieurs orthographes différentes ; de même, l'orthographe adoptée pour un certain nombre de mots, notamment ceux contenant une voyelle longue *ā*, que les graphies coraniques archaïques pouvaient ignorer, sont systématiquement régularisées conformément à l'orthographe arabe standard ;

– d'autre part, le texte électronique, quoiqu'intégralement vocalisé, ne comporte pas les signes diacritiques à visée spécifiquement orthoépique, en particulier ceux qui réglementent les phénomènes d'assimilation phonétique et de pause dans la lecture coranique traditionnelle[12].

2. Quelques caractéristiques statistiques du texte coranique

L'analyse[13] d'un corpus électronique du texte coranique tel qu'il vient d'être caractérisé livre les résultats statistiques résumés dans le tableau suivant :

Nombre de formes	17 503
Nombre d'occurrences	77 874
Rapport occurrences/formes	4 449
Hapax legomena	10 899
Hapax dislegomena	2 694

Les « occurrences » sont les mots graphiques (séquence de lettres séparées par deux « séparateurs »). En arabe, langue agglutinante, il peut s'agir de mots sim-

12. Sur ces signes dans le « Coran de Fouad », voir Humbert (1980).
13. À l'aide du logiciel d'analyse de Corpus TACT, Lancashire *et al.* (1996).

ples ou de mots complexes[14] constitués par un mot simple auquel s'agglutinent des prépositions ou conjonctions clitiques d'une part, des pronoms clitiques d'autre part. Il y en a 77 874 dans le Coran analysé (à comparer avec les résultats de « la docte Sympathie »).

Les « formes » sont, en quelque sorte, les « modèles » des occurrences, décomptés une seule fois, en sorte qu'à toutes les occurrences identiques est associée une seule forme. Le nombre des formes, ici 17 503, donne une idée du vocabulaire total mobilisé par le texte. Mais cette idée est imprécise et en l'occurrence très surévaluée compte tenu de ce qui vient d'être dit sur la notion d'occurrence. Par exemple, les trois vocalisations syntaxiques d'un même nom seront comptabilisées comme trois formes différentes. De même, toutes les formes cliticisées d'un même mot. On comprend alors en quoi le nombre des formes est une très grosse surévaluation du lexique réel d'un texte, en tout cas en arabe.

Cependant, le rapport occurrences/formes donne une certaine idée de la richesse lexicale d'un texte, car tendanciellement, plus ce rapport est élevé, plus le taux de répétition d'une même forme est grand, et donc moins le texte est varié. En l'occurrence, la valeur de ce rapport pour le Coran, proche de 4,5, est relativement élevée, suggérant un assez fort taux de répétition. Nous verrons plus loin que d'autres indices confortent cette prédiction.

Les « hapax legomena » sont les mots graphiques qui n'apparaissent qu'une seule fois dans le texte, ou si l'on veut, les cas ou l'occurrence est aussi la forme. En principe, plus le taux d'hapax legomena est élevé, moins il y a de répétitions et plus le texte est riche lexicalement. Le nombre d'hapax legomena dans le Coran est loin d'être négligeable (près de 11 000 formes sur 17 503). Mais cette valeur, rapprochée de celles qui précèdent, suggère paradoxalement que les formes restantes présentent un taux de répétition d'autant plus élevé. En somme, le Coran présente près de 11 000 formes qui n'apparaissent qu'une fois dans le texte, et environ 6 500 formes ayant un taux de répétition égal ou supérieur à 2. Les « hapax dislegomena » sont précisément les mots graphiques qui n'apparaissent que deux fois dans le texte.

L'examen du texte à l'aide d'un logiciel de traitement de corpus permet également d'établir la liste des formes classée soit par ordre alphabétique, ce qui est banal, soit par ordre fréquentiel, ce qui est beaucoup plus éclairant, et peut notamment servir de point de départ pour une exploration thématique du texte.

Il ne saurait, bien sûr, être question de reproduire ici de telles listes, puisque les 17 503 formes du texte coranique, même imprimées à 100 lignes par page (ce qui ne serait guère lisible), exigeraient 175 pages ! Nous nous contenterons donc, pour en donner une petite idée, de reproduire les 37 premières lignes de la liste fréquentielle du texte coranique, et de faire dessus quelques commentaires généraux.

14. Sur cette notion et la structure des mots complexes, voir Kouloughli, 1994, chapitre 1.

Formes (ordre lexicométrique)	Fréquence
min	2 363
fiy	1 207
maa	1056
ellaahi	828
laa	816
el^{x}a@iyna	810
ellaahu	734
&alaY	669
’il^{x}aa	662
walaa	658
wamaa	645
’in^{x}a	607
ellaaha	591
’an	539
qaala	416
’ilaY	404
man	402
lahum	373
’in	357
yaa	350
cum^{x}a	337
lakum	336
bihi	327
kaana	323
bimaa	296
qul	293
’aw	280
@aalika	280
lahu	275
el^{x}a@iy	268
huwa	265
hum	260
’aamanuwe	253
qaaluwe	250
&an	244
fiyhaa	240
waman	240

Coran : liste des 37 premières formes les plus fréquentes

L'examen de la liste fréquentielle[15] permet, entre autres, les observations suivantes :

– les formes les plus fréquentes sont, typiquement, des « mots grammaticaux » (prépositions, conjonctions, relatifs, etc.). Cette propriété se retrouve dans tout texte suffisamment long quel qu'en soit le sujet ou la langue ;

– par contre, une caractéristique tout à fait frappante de la liste de formes les plus fréquentes du texte coranique est que le mot *Allāh* (transcrit ici « *ellaah* ») apparaît avec une fréquence exceptionnellement élevée puisque, sous ses trois formes fléchies, il totalise 2 153 occurrences[16], ce qui lui donne une fréquence d'occurrence absolument unique pour une unité lexicale « pleine » ;

– d'autres « mots pleins », noms ou verbes, apparaissent avec une fréquence significativement élevée dans cette liste, comme par exemple les noms *ʾarḍ* (terre) ou *samāwāt* (cieux) ou les verbes *ʾāmana* (croire) ou *kafara* (mécroire).

On voit certes, d'après leur sémantisme, que les « mots pleins » à fréquence d'occurrence élevée sont fortement « thématiques » dans le texte coranique, et que ceci peut expliquer cela. Mais nous allons voir, en explorant un autre paramètre statistique du texte, celui des « segments répétés », que cette fréquence élevée peut en partie s'expliquer par une autre propriété du texte coranique : son taux élevé de répétitions.

On entend par « segment répété » (désormais SR) toute suite de formes (au sens défini plus haut) dont la fréquence est égale ou supérieure à 2 dans un texte. Le logiciel Lexico3 est doté d'une fonctionnalité permettant de repérer et de comptabiliser tous les segments répétés d'un texte. L'exécution de cette fonctionnalité sur le corpus coranique révèle des propriétés assez remarquables de ce point de vue. En effet, à côté de SR que l'on peut considérer comme correspondant simplement à des (fragments de) structures syntagmatiques de la langue comme par exemple *min baʿdi* (« après », 80 occurrences) ou *ʿalā kulli* (« sur tout », 55 occurrences), on trouve des SR qui sont incontestablement liés au contenu thématique du texte comme *allaḏīna ʾāmanū* (« ceux qui ont cru », 184 occurrences) ou *walahum ʿaḏābun ʾalīmun* (« et pour eux un châtiment douloureux », 12 occurrences).

Ces SR sont de longueur variable, constitués de groupes pouvant aller jusqu'à 11 mots comme, par exemple, *ʾillā allaḏīna tābū min baʿdi ḏālika waʾaṣlaḥū faʾinna allāha gafūrun raḥīmun* (« sauf ceux qui se repentiront ensuite et se réformeront car en effet Dieu est Pardonneur et Miséricordieux », 2 occurrences). Les SR les plus longs (longueur comprise entre 8 et 11) n'ont en général que des fré-

15. Cette liste a été produite par le logiciel de traitement de corpus Lexico3 (André Salem *et al.*), à partir d'une version du texte coranique ayant subi une transcription « phonographématique », c'est-à-dire une transcription qui respecte les conventions orthographiques de l'arabe. La lecture du texte transcrit ne devrait pas s'avérer trop difficile même pour le lecteur qui ne connaît pas le détail du système.

16. Auxquelles il faudrait ajouter les formes cliticisées comme *waellaahu* ou *biellaahi*, formes très fréquentes et qui viennent encore grossir le nombre total d'occurrences du nom divin dans le Coran.

quences d’occurrence modestes[17] (généralement comprises entre 2 et 4). Mais à partir de la longueur 7 on peut trouver des fréquences de 9, et à la longueur 6 la fréquence peut approcher la vingtaine ! Et surtout, le nombre de ces SR est très important : souvent plusieurs dizaines pour chaque ordre de longueur !

Ces propriétés du texte coranique que constituent le nombre, la longueur et la fréquence d’occurrence de SR est sans doute à rapporter à son statut originel de texte oral. En effet, nombre de ces SR sont, *mutatis mutandis*, les équivalents des « formules » que les chercheurs ont mises en évidence dans la poésie ancienne[18]. Cette propriété du texte coranique n’a en tout cas pas échappé aux commentateurs anciens et modernes qui lui ont parfois consacré des études spécifiques[19].

3. Quelques caractéristiques lexicographiques du texte coranique

Nous terminerons cette brève présentation de données lexico-statistiques relatives au texte coranique par un examen plus spécifique de son lexique.

On sait que, dans le cas général, tout mot arabe s’analyse en une racine de trois radicales consonantiques ou plus et en un schème constitué de voyelles et éventuellement de consonnes se préfixant, s’infixant ou se suffixant sur les radicales constituant la racine.

On peut souhaiter, pour les besoins de la procédure de traitement statistique, faire en sorte que toute unité lexicale, même si elle n’est pas *stricto sensu* analysable en racine et schème (c’est le cas des « particules » de la grammaire arabe traditionnelle), puisse se voir associer une « pseudo-racine » conventionnelle. Cette pseudo-racine pourrait être constituée des consonnes qui forment l’unité considérée : ainsi, la particule *fī*, graphiée « فِي » et transcrite phonographématiquement *fiy*, se verrait associer la « pseudo-racine » conventionnelle <*fy*>. Si en outre on poussait un peu plus loin ce principe, on conviendrait que les « lettres mystérieuses », qui occurrent à l’initiale de certaines sourates, et qui sont généralement décomptées comme constituant en elles-mêmes un verset coranique, que ces lettres donc ont le statut de formes du lexique coranique, et doivent à ce titre se voir associer une « pseudo-racine » conventionnelle. Ainsi, la lettre « ق » qui occurre comme la première forme du premier verset de la sourate du même nom (sourate 50) se verra associer la pseudo-racine <*q*>. Évidemment cela conduit à admettre l’existence de pseudo-racines biconsonantiques et même monoconsonantiques.

Si l’on adopte les conventions ci-dessus et que l’on se livre, sur cette base, à l’analyse morphologique de l’ensemble du corpus coranique, on aboutit aux ré-

17. Mais un SR de longueur 10 a, dans le Coran, une fréquence de 6 ! Il s’agit du SR *qāla yā qawmī ʿbudū llāha mā lakum min ʾilāhin ġayruhu* (« Ô mon peuple adorez Dieu, vous n’avez pas d’autre dieu que lui »).
18. Voir par exemple Monroe (1972).
19. Par exemple, Muḥammad Ibn Ḥamza al-Kirmānī (m. 1111), *ʾAsrār al-takrār fī l-Qurʾān*.

sultats suivants : le corpus coranique compte 1 767 racines et pseudo-racines ainsi ventilées :

Pseudo-racines monoconsonantiques	4
Pseudo-racines biconsonantiques	33
Racines (ou pseudo-racines[20]) triconsonantiques	1 637
Racines (ou pseudo-racines) quadriconsonantiques	65
Racines (ou pseudo-racines) pentaconsonantiques	22
Racines (ou pseudo-racines) hexaconsonantiques	3
Racines (ou pseudo-racines) de 7 consonnes[21]	3
Racines (ou pseudo-racines) de longueur supérieure à 7	0

En ce qui concerne le nombre de formes auxquelles ces racines[22] donnent naissance, voici quelques valeurs significatives :

503 racines ne sont associées qu'à une seule forme.
245 racines ne fournissent que 2 formes.
145 racines ne fournissent que 3 formes.
100 racines fournissent 4 formes.
101 racines fournissent 5 formes.
57 racines fournissent 6 formes.
52 racines fournissent 7 formes.
47 racines fournissent 8 formes.
51 racines fournissent 9 formes.
27 racines fournissent 10 formes.
...
15 racines fournissent plus de 100 formes.

La racine <'*ty*>, la plus productive, fournit 227 formes.
En ce qui concerne le nombre d'occurrences : 413 racines ne fournissent qu'une seule occurrence (et bien sûr, qu'une seule forme).

213 racines ne fournissent que 2 occurrences (de 1 ou 2 formes).
125 racines fournissent 3 occurrences (de 1 à 3 formes).
95 racines fournissent 4 occurrences (de 1 à 4 formes).
89 racines fournissent 5 occurrences (de 1 à 5 formes).
55 racines fournissent 6 occurrences (de 1 à 6 formes).
36 racines fournissent 7 occurrences (de 1 à 7 formes).
29 racines fournissent 8 occurrences (de 1 à 8 formes).
39 racines fournissent 9 occurrences (de 1 à 9 formes).
29 racines fournissent 10 occurrences (de 1 à 10 formes).
...
10 racines fournissent plus de 1000 occurrences.

La racine <*qwl*> fournit 1 722 occurrences pour 120 formes.

---

20. Il faut prévoir l'existence de pseudo-racines de trois consonnes ou plus car certaines séquences de « lettres mystérieuses » ont des longueurs supérieures à 2 : c'est le cas par exemple du premier verset de la sourate 2 (longueur 3) ou de celui de la sourate 7 (longueur 4).
21. Il s'agit, on l'aura peut-être deviné, des pseudo-racines associées aux noms propres *ʾibrāhīm*, *ʾisrāʾīl* et *ʾismāʿīl*.
22. Ou pseudo-racines. Cette précision sera sous-entendue dans la suite.

La racine (ou pseudo-racine) la plus riche en occurrences est <*’n*>, associée aux formes nombreuses et diverses des particules *’inna*, *’anna*, *’an*, *’in* et qui fournit 4 037 occurrences. Elle est suivie de la pseudo-racine <*mn*>, associée aux formes *min* et *man* et qui fournit 3 099 occurrences. Vient ensuite la racine <*’lh*>, associée aux diverses formes du nom divin et qui fournit 2 851 occurrences.

Signalons en passant que sur les 1 767 racines et assimilées décomptées dans le Coran, 1 200 fournissent des formes verbales. Sur ces 1 200 racines, 15 sont quadriconsonantiques, et toutes les autres triconsonantiques[23].

Éléments de bibliographie

Sources anciennes

Al-Qur’ān al-karīm, Le Caire, 1923.

’Abū ‘Amr al-Dānī (m. 1052), *Al-Muqni‘ fī rasmi maṣāḥifi al-’amṣāri*, Le Caire, Maktabat al-Kulliyyāt al-’azhariyya, 1978.

Ibn Muǧāhid (m. 936), *Kitāb al-sab‘a fī l-qirā’āt*, Le Caire, Dār al-Ma‘ārif, 1979.

Muḥammad Ibn Ḥamza al-Kirmānī (m. 1111), *’Asrār al-takrār fī l-Qur’ān*, Le Caire, Dār al-faḍīla, 1977.

Ǧalāl al-Dīn al-Suyūṭī (m. 1505), *’Asrār tartīb al-Qur’ān*, Le Caire, al-Maktaba al-‘aṣriyya li-l-ṭibā‘a wa-l-našr, 2003.

Ibn Šihāb al-Zuhrī (m. 741), *Kitābu tanzīli l-Qur’ān*, Beyrouth, Dār al-kitāb al-jadīd, 1980.

Sources modernes

Blachère R., 1958, *Introduction au Coran*, Paris, Maisonneuve et Larose.

Chouémi M., 1966, *Le verbe dans le Coran*, Paris, Klincksieck.

Déroche F., 2005, *Le Coran*, Paris, PUF (Que sais-je ?).

Humbert G., 1980, « Essai d’interprétation linguistique des conventions orthographiques du Coran (édition égyptienne) », mémoire de maîtrise (non publié), Université Paris 8.

Kouloughli D. E., 1982, *Sur la phonographématique arabe*, *Analyses/théorie,* n° 1, p. 79-151.

— 1994, *Grammaire de l’arabe d’aujourd’hui*, Paris, Presse Pocket.

Lancashire I. *et al.*, 1996, *Using TACT with electronic texts*, New York, The Modern Language Association of America.

Mardrus J.-C. (traducteur), 1899, *Les Mille et une Nuits*, Paris, Laffont.

Monroe J. T., 1972, « Oral composition in pre-islamic poetry », *Journal of Arabic Literature*, vol. III, p. 1-54.

Nöldeke T. et Schwally F., 1909/1919/1926, *Geschichte des Qorans*, Leipzig, Dieterich'sche Verlagsbuchhandlung. Traduction arabe : *Tārīḫ al-Qur’ān*, 2004, Beyrouth, Konrad-Adenauer Stiftung.

Salem A. *et al.*, 2005, Lexico3, Équipe CLA2T, Université Paris 3.

23. Pour une étude détaillée des racines verbales du Coran, voir Chouémi (1966).